Incorporating Uncertainty Metrics into a General-Purpose Data Integration System

Brenton Louie[1], Landon Detwiler[3], Nilesh Dalvi[4], Ron Shaker[2], Peter Tarczy-Hornoch[1,2,4], and Dan Suciu[4]
University of Washington Departments of Medical Education and BioMedical Informatics[1], Pediatrics[2], Biological Structure[3], and Computer Science and Engineering[4]
brlouie@u.washington.edu, det@u.washington.edu, nilesh@cs.washington.edu, rshaker@u.washington.edu, pth@u.washington.edu, suciu@cs.washington.edul

Abstract

There is a significant need for data integration capabilities in the scientific domain, which has manifested itself as products in the commercial world as well as academia. However, in our experiences in dealing with biological data it has become apparent to us that existing data integration products do not handle uncertainties in the data very well. This leads to systems that often produce an explosion of less relevant answers which subsequently leads to a loss of more relevant answers by overloading the user. How to incorporate functionality into data integration systems to properly handle uncertainties and make results more useful has become an important research question.

In this paper we describe an enhanced general-purpose data integration system which incorporates uncertainty metrics within a formal probabilistic framework. Additionally, for evaluation purposes, we have implemented a use case scenario which utilizes biological data sources and performed a study which provides validation of system query results.

1. Introduction

Unlike *more* "traditional" sources of data such as banking or inventory, biomedical scientific data is inherently uncertain. This uncertain nature presents unique challenges in regards to its handling and manipulation, "information overload" being one example [1]. For us in particular, data uncertainties have made it difficult to apply data-integration (DI) technologies according to the wishes of our biological collaborators. This challenge has led us down our current path, which is the incorporation of a formal framework for handling data uncertainty (uncertainty metrics) into a general-purpose federated DI system. To help illustrate uncertainty metrics and how they are applied to biological data, we provide examples of data uncertainty drawn from our own experiences. Although what we provide are examples, we feel that they capture and classify general areas of uncertainty in biological data into two broad categories:

1) *Inherent Data Uncertainties*. Inherent data uncertainties are attributes of the data itself and not artifacts of its representation. Data generated from laboratory experimental methods often have inherent uncertainties. To illustrate an extreme case, two-hybrid screening assays, which are used to detect protein interactions, have error rates estimated to be close to 50% [2]. Experimental data can also be generated from computational (or "in-silico") experiments. The BLAST algorithm [3] searches in a database for sequences similar to a query sequence. The similarity between any two sequences is measured by the BLAST "e-value", which is the degree to which the pairing could occur by chance. Additionally, uncertainties can be rooted in the ever-evolving nature of biological knowledge itself. For example, GenBank references sequences (RefSeq's) are assigned "status codes" which refer to the amount of evidence and expert curation attributed to a given sequence and its function [4]. These codes range from "inferred" where there is little support for a given sequence, to "reviewed" where substantial evidence exists and has been vetted by a biological domain expert. Status codes for sequences change over time as evidence for them accumulates.
2) *Data Representation Uncertainties*. Data representation uncertainties result from the

mapping of real world information onto a computable representation of this information. At last count there were over 600 online data sources in molecular biology [5]. Unfortunately, for all the data that is available there are no common standards for representing it (in part due to the evolving nature of biomedical knowledge). The result is the decentralized and heterogeneous nature of biological data sources which is an underlying source of many data uncertainties. For instance, there is no common identifier for a biological object [6] which make it difficult to query across data sources (manually or otherwise), a task which is commonly performed. Linkages between data records may then require string matches on text fields rather than more reliable "foreign-key" relationships. Additionally, data sources tend to represent data in idiosyncratic fashion. For example, GenBank uses RefSeq status codes to represent the level of evidence for a particular gene but the Gene Ontology (GO) uses evidence codes [7]. Given evidence from both sources, it is sometimes difficult to make comparisons, such as determining which code provides the greater weight of evidence.

1.1. Functional Annotation of Proteins

These uncertainties in biological data pose significant challenges to existing DI systems. In conjunction with our biological collaborators, we applied DI technology to the task of assigning biochemical function to unknown protein sequences (functional annotation). Functional annotation is crucial to bioresearch and is fast becoming a high priority in the biological community due to the vast numbers of new proteins predicted from genome sequencing projects which require function assignment [8, 9]. Unfortunately, annotation is currently quite expensive in terms of time and human capital as it is a very manual process which requires a great deal of expertise due to its complex and often ambiguous nature. Functional annotation is also highly dependant on various biological data sources which are developed in isolation and are thus very heterogeneous and dispersed. One of the many challenges faced by biological researchers is querying, traversing and compiling data from a multitude of different sources (Figure 1).

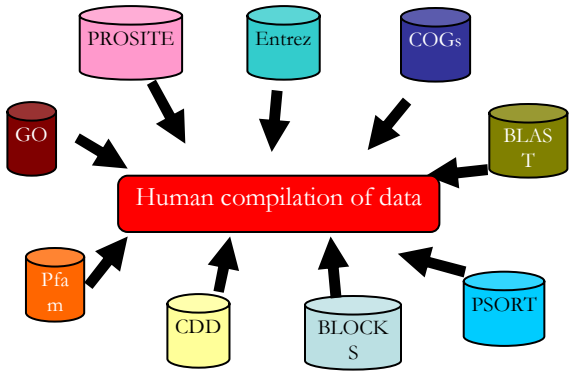

Figure 1. In the current paradigm for annotating proteins, biological researchers often compile data from various non-interoperable sources manually (or use ad-hoc techniques) which greatly slows the pace of research (adapted and modified from [10]).

To address this issue and drawing from our experience in integrating data sources for functional annotation, we were able to properly integrate a number of data sources using an existing data integration system (BioMediator) with the intent of creating an easily accessible annotation data resource for biological researchers but uncertainties in the data greatly affected the usability of the system by producing *explosions* of less relevant answers to queries. This overwhelmed human users making it hard to find the most relevant answers. This problem is rooted in the fact that protein annotation is knowledge-base driven. What this means is that computational function assignment is dependant upon determining the "similarity" between a protein of unknown function and previously characterized proteins. These similarity or classification functions are inherently probabilistic in nature, Hidden Markov Models for determining protein family classification being one example (Pfam [11]). In many cases, multiple results are returned from each individual source and it can be difficult, for a data integration system as well as a human, to determine what the best result is from a particular source or among several sources (Figure 2). Generally speaking, however, the result with the strongest overall similarity score deserves the most attention as it is likely the most probable annotation for a given protein, given the caveats that similarity scores may not be totally comparable between sources. We determined that a possible approach to addressing the issue of winnowing out good results from a data integration perspective would be to include functionality in the DI system

which would allow it to handle uncertainties in the data. This would enable it to perform functions such as highlighting only the most relevant data by presenting results as a ranked list.

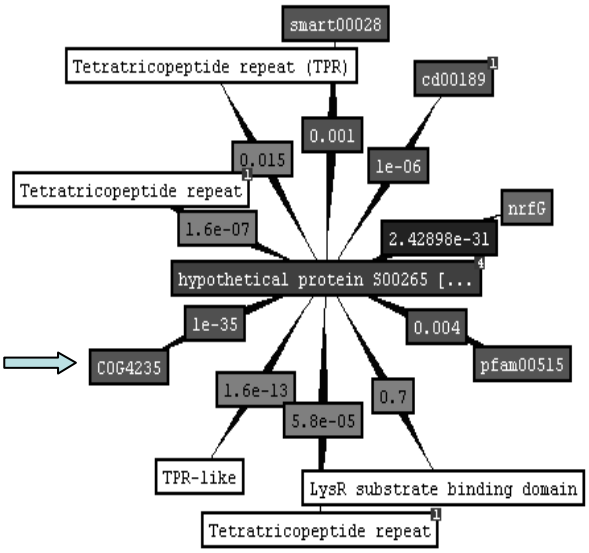

Figure 2. Simplified example of an integrated result set of annotation data output from the BioMediator system when queried with *S. oneidensis* reference protein SO0265 (center of graph). There are many related results from multiple individual data sources with results such as "Tetratricopeptide repeat" or "nrfG" (a gene name). The result which agrees with the reference annotation however is from COG4235 (arrow) the description of which is "Cytochrome c biogenesis factor". This disagrees with many of the results from the other sources and it would be difficult for an annotator to select this result from all the others without some mechanism to highlight more relevant results (based on common metrics.

In this paper we discuss our approach to incorporating a formal framework for handling data uncertainties into a general-purpose data integration system. In addition we have carried out an initial study which validates that our system can produce ranked lists of results which are meaningful to a biological domain expert.

2. Data Model

2.1. The BioMediator Data Integration System

The BioMediator system, developed at the University of Washington, was designed as a tool for integrating syntactically and semantically heterogenous data while maintaining data source autonomy. Although it has been particularly applied in the biological domain, the system is general-purpose [12-18] and freely available (www.biomediator.org).

2.1.1. Background: BioMediator Architecture. BioMediator is a mediated schema distributed data integration system. It provides a unified, uniform view over a network of inter-related data sources. Source object types (e.g. EntrezGene_Gene and SwissProt_Protein) are aligned across sources by mapping them to common object types within the mediated schema (e.g. Gene and Protein respectively). A query result object (i.e. a specific Gene record) returned by the system is uniquely identified by its mediated schema type, data source, and source uid.

BioMediator exposes query result sets using an annotated directed graph data model. Nodes in this graph represent source records, returned during the query process, transformed into objects in the mediated schema space. Nodes are annotated with their appropriate attribute values (such as schema object type, source, and uid). Edges in the graph represent cross-references from one node (result object) to another. Such references may be explicitly represented in the head node ("foreign key" references) or they may be derived by the system (see Query Model below).

2.1.2. Query Model in BioMediator. BioMediator employs an exploratory, browser-based query model. Users first initiate a *seed query*, which is constructed by specifying the desired object type from the mediated schema and desired attribute-value constraints (e.g. Gene:symbol='BRCA1'). The system queries all sources with mappings to the corresponding object type and retrieves from them records satisfying the given constraints. The initial result graph is represented as a set of result nodes linked from a system node representing the seed query (see Figure 3). Through a process called *query expansion*, users select any subset of nodes from the current result graph for further exploration. The system attempts to join the selected nodes with related records from the source network. This process may include following foreign key references from a selected node, querying sources in the network that might themselves have foreign key references back to a selected node, or performing inter-source record alignment by matching values found in non-key fields. Depending on the sources involved, these latter types of alignments may employ partial matching algorithms.

While the set of nodes to be expanded can be explicitly selected by the user, there is also an "*expand all"* option. As the name implies, *expand all* triggers a query expansion over the set of all current result nodes. As new result nodes are added to the result graph, new edges are also added to illustrate the inter-node referential links. Figure 4 shows a portion of a result graph after performing an *expand all* operation on the seed result graph from Figure 3. This figure visually suggests the data overload problem that can ensure from an exploratory graph model. For more details on our browser-based query model, please see [13].

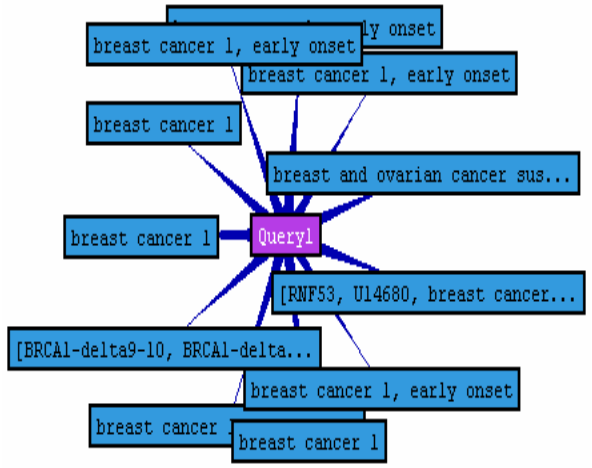

Figure 3. Seed query (BRCA1 gene name, center of graph) and initial results after one *expand all* operation.

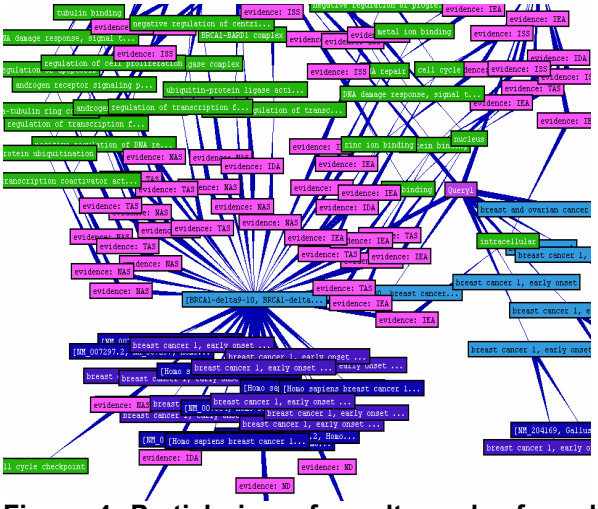

Figure 4. Partial view of result graph of seed query (BRCA1 gene name) showing results from multiple sources after 3 *expand all* operations. The size of the result set quickly becomes difficult for a human to analyze.

2.2. Data Uncertainties in BioMediator

2.2.1. Uncertainty Extensions. In this work, we have embarked on a new approach by extending the BioMediator data and query models to better support the handling of uncertainty in the data integration process for the overall purpose of making result sets more useful.

2.2.2. Uncertainty Metrics. Uncertainty metrics provide a framework in BioMediator for describing such things as the quality of data sources, quality of cross-references (links) between sources, and the quality of the data entities themselves and links between them. The uncertainty metrics are interpreted probabilistically and are stored as annotations on the result graph. The system then calculates relevance scores for each query result which allow results to be ranked, enabling users to identify the most useful results. The following is a summary description of the four fundamental uncertainty metrics that capture all types of uncertainty in the BioMediator system:

1) *Ps measure*: A quantification of a user's prior belief in the quality of data records of a particular mediated schema type from a particular data source (e.g. *Genes* from Entrez Gene, *Classifications* from Entrez Gene, and *Classifications* from GO, are each assigned Ps values). Ps sets the maximum belief threshold for any particular data record of the corresponding type and source. For example, consider the comparison between proteins from SwissProt and TrEMBL. SwissProt is a manually and carefully curated data source of protein functional information whereas TrEMBL contains only computational predictions which are deemed less reliable. The class of protein records from SwissProt therefore should be assigned a higher Ps value than those from TrEMBL because SwissProt protein records are generally trusted to a greater degree.
2) *Qs measure*: A quantification of a user's prior belief in the quality of a particular relationship type from the mediated schema where a relationship type asserts the schema type and source of both the head and tail records (e.g. *Genes* in Entrez Gene to *Proteins* in Entrez Protein). Qs sets the maximum belief threshold for any particular data link whose head and tail records are of the corresponding types and sources. To elaborate, records in some sources, such as *Gene* records from Entrez Gene, contain

references to records in another source, such as *Protein* records from Entrez Protein. In this example, these references are in the form of "accession" numbers which essentially correspond to unique identifiers (foreign-keys). Records between other types and sources however may only be connected by non-foreign keys, e.g. text-string similarities such as is the case between *Genes* from Entrez Gene and *Genes* in OMIM. In this example, the relationship between *Genes* in Entrez Gene and *Proteins* in Entrez Protein should be assigned a higher Qs value since these links are better in general than those between *Genes* in Entrez Gene and *Genes* in OMIM.

3) *Pr measure*: This is a quantification of a user's belief in a particular data record. Unlike the Ps measure, the Pr measure is calculated at the time a particular result is returned from a particular source. It is used to capture data uncertainties which differ between records of the same type and source. Gene records in Entrez Gene, for example, are attributed with a Refseq status code which ranges in value from "inferred" to "reviewed". These status codes correspond to the amount of evidence for a given gene, therefore "reviewed" should be assigned a higher Pr value than "inferred".
4) *Qr measure*: This is a quantification of a user's belief in a particular cross-reference (link) between two data records. Like Pr, it is dynamic (calculated at the time two linked results are returned by the system). For example, record cross-references using unique identifiers always receive a Qr of 1.0. Some records may reference others via the use of comparison algorithms such as BLAST. For BLAST cross-references Qr's are dynamically computed by converting the e-value from the BLAST algorithm into a numeric value between 0.0 and 1.0. BLAST comparisons which correspond to better "matches" between records (higher similarity) receive higher Qr values.

2.2. Probabilistic Query Evaluation

The uncertainty metrics discussed thus far have all been local measures. They are calculated and applied to nodes and edges irrespective of their position in the greater result graph. The information we are seeking, however, is a global measure of the relevance of each node to the original seed query. Such a relevance measure should be based on the uncertainties of nodes and edges along all paths connecting the seed query node to the target node. To help us solve this problem, we recast it in terms of a network reliability problem [19]: For each result node n_1, $Ps_{n1} * Pr_{n1}$ is interpreted as the probability that the node is currently present in the network. Likewise $Qs_{e1} * Qr_{e1}$ is viewed as the probability that a given network link e_1, is available. We calculate the relevance of a node as the probability that the node is reachable from the seed node in our network reliability problem.

2.3. Relevance Evaluation Algorithm

2.3.1. Relevance score calculation. The problem is as such: given a directed graph with probabilistic scores on nodes and edges, compute for each node the probability that there is a path from the start (seed) node to any given result node. Computation of exact probability scores is intractable but can be efficiently approximated to arbitrary precision using simulation algorithms [20]. In our case, we simulate in a single pass *N* trials (path traversals) where nodes and edges are included in the traversal with associated probabilities. This is performed by storing a randomized N-bit trial vector associated with each node and edge, where each bit is a binary value denoting success or failure of a particular trial (based on the uncertainty metrics). In a depth first search of the graph beginning from the seed node, we populate a success vector for each node indicating for each trial whether or not that node is reachable by some path (i.e. the path contains only nodes and edges included in this trial). For each node a count (*k*) is maintained which is the number of times it could be reached, via some path, over the total number of trials. The final score, or relevance, for a node is then estimated using the quantity *k/N*, where *k* is the number of set bits in a node's success vector. Because this is an approximation algorithm, the choice of *N* influences the error in the estimation and the larger the *N*, the smaller the error. Also, for any fixed value of *N*, the greater the actual relevance of a particular node, the better the approximation will be. Overall, the algorithm should correctly rank the most relevant answers (which are the most important), whereas the poorest results may be slightly out of order.

2.3.2. Relevance Score Measure. The final relevance score of a node is simply the probability that the node can be reached from the seed node, e.g. the probability that there is a path from the seed node to any particular result node. Figure 5 illustrates an example of a result

with various nodes annotated with uncertainty metrics and relevance (UII) scores. Relevance scores are useful in that entity sets can be sorted accordingly and results presented to users as ranked lists.

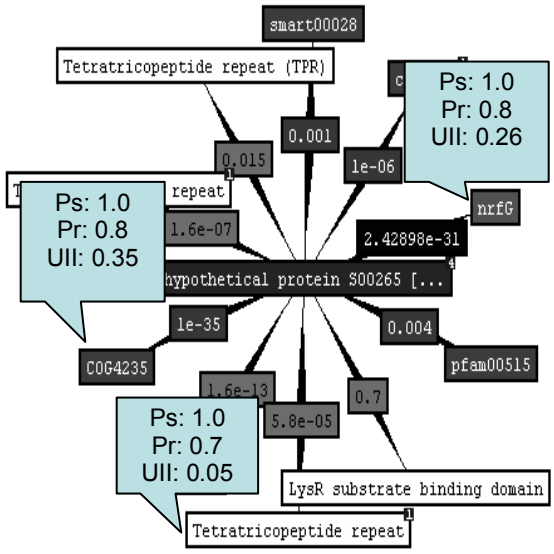

Figure 5. Same result graph as in Figure 2 for *S. oneidensis* protein SO0265 (center of graph) with some results annotated with examples of uncertainty metrics (Ps, Pr) and final relevance scores (UII). The highest relevance score generated by the system is from COG4235 (description agrees with the reference annotation) because in this example, COG4235 has the best quality path from the query node. The relevance score can be utilized by the system to rank, highlight, or filter in order to enable annotators to select the best results.

3. Experimental Results

3.1. Experimental Setup

3.1.1. Use Case. To validate query results we implemented a simple functional annotation use case based on the classification of proteins from the Clusters of Orthologous Groups (COG) database [21]. The COG database organizes known proteins into groups according to general biochemical function (COGs) based on sequence similarity. Biologists often assign a biochemical function to an unknown protein by submitting it as a query to the COG database which subsequently compares the sequence of the unknown protein to the various COGs and returns results (often multiple). Each result carries with it biochemical function information about the COG as well as a probabilistic similarity measure (e-value) of the COG to the query protein. The rationale here is that given a protein with an already known COG assignment, the system should be able to take that protein as a query and reproduce the actual COG assignment by ranking it at the top of a results list. Therefore, a set of 32 proteins with known and pre-defined biochemical functions were chosen according to the 25 functional categories in the COG database (e.g. transcription, cell cycle, etc.) with the help of a collaborating biologist. These served as the "gold-standard" by which the system queries were evaluated.

3.1.2. BioMediator. An initial set of functional annotation data sources were identified by our biological collaborators. There are: AmiGO [22], the Conserved Domain Database (CDD [23]), Entrez Gene and Entrez Protein [24], Pfam [25], SuperFamily [26], the BLAST database at the NCBI, the PSI-BLAST database at the DDBJ, PDB [27], and UniProt [28]. Note that some of the data sources are themselves aggregates of sources. For instance, CDD contains Pfam, SMART [29], COG, as well as its own proprietary information. UniProt contains SwissProt, and TrEMBL [30]. Appropriate data entities, attributes, and their relationships were identified and modeled in the mediated schema. It was deemed important to test queries on multiple data sources to ensure that: 1) all uncertainty metrics were populated correctly, and 2) relevance scores were calculated correctly for all data entities. It would be a better test of the system, for example, if COG rankings were evaluated using result sets which included data from multiple sources rather than just the COG.

3.1.3. Population of Uncertainty Metrics. All entities from sources in the federation were considered to be of equivalent quality, thus Ps values for all were set to 1.0. Qs values were also all set to 1.0 since links between all sources were either "foreign-key" cross-references or comparison algorithms (which are handled by the Qr metrics). Other metrics were determined from data type descriptions of individual data sources (e.g. some data sources explicitly say which data to trust more) and also in discussions with domain experts as to their opinion of what data was most relevant to them for the particular task of annotating proteins. These initial uncertainty metrics were meant to serve as prototypical examples with which to test the calculation of relevance scores by the system and provide generally meaningful results for evaluation purposes. Tables 1 and 2 provide selected Pr and Qr values incorporated in the system, not all the metrics were evaluated in this study but are provided here for the purposes of illustration:

Table 1. Pr belief values for entity types. Attempts were made to assign identical Pr values to entity attributes with similar meaning such "TAS" (Traceable Author Statement) and "Reviewed" which both suggest some involvement by a human annotator.

Entity	Attribute	Calculation
Gene	Status Code	Reviewed (1.0) Validated (0.8) Provisional (0.7) Predicted (0.4) Model (0.3) Inferred (0.2)
Classification	Evidence Code	IDA (1.0) TAS (1.0) IGI (0.9) IMP (0.9) IPI (0.9) IEP (0.7) ISS (0.7) RCA (0.7) IC (0.6) NAS (0.5) IEA (0.3) ND (0.2) NR (0.2)
Domain, COG	Description	Curated (1.0) Alignment from Source (0.5)

Table 2. Qr belief calculations for selected entity relationship types. There is also a range check which assigns a Qr of 0.0 or 1.0 if the result of the calculation is less than 0.0 or greater than 1.0 respectively.

Relationship	Attribute	Calculation
Protein -> Domain Protein -> COG Protein -> Protein	e-value expect	$abs\left(\frac{\log_{10}(evalue)}{200}\right)$

3.1.4. Evaluation. The 32 gold-standard proteins were submitted to the system which subsequently queried all databases in its federation (including the COG), populated belief values for all data entities, calculated relevance scores, and returned results as ranked lists. The functional category of the highest ranking COG in a result set from the system was then compared to the actual COG functional category from the gold-standard protein. The evaluation methodology is a simple percent agreement between the top ranking COG in the system versus the COG category from the gold-standard protein.

3.2. Results

The 32 gold-standard proteins with pre-assigned COG functional categories were submitted as seed queries to the system and subjected to query expansion. All data sources in the federation were queried successfully, belief values for all data entities populated correctly, relevance scores calculated, and results returned as ranked lists of entities. Of the 32 seed proteins, 14 had only a single result from the COG. These were omitted from the final analysis because no ranking was necessary (although this did mean that the system produced a correct answer by default). The top COG result from each of the remaining 18 proteins was selected, the COG functional category determined and subsequently compared with the gold-standard COG category of the protein. The initial agreement between the COG category from the system and the gold-standard COG category was 77.8% (14/18). However, upon further inspection it was discovered that three of the gold-standard proteins were actually assigned to two COG functional categories and the system had actually correctly ranked at least one of them. These three results were then converted to successes which increased the agreement to 94.4% (17/18). The remaining disagreement was inspected and it was determined that the reason for the incorrect result was due to the approximating nature of the relevance score calculation. After adjusting a parameter in the algorithm (increasing the number of trials) and re-submitting the query, the top-ranking COG in the system turned out to be the correct one, bringing the total agreement to 100% (18/18).

4. Observations

The study results, although preliminary, are encouraging and strongly suggest that our formal framework is performing adequately by quantifying the quality of result sets and ranking them according to a biologically significant metric. In regards to the single case where the system produced an incorrect ranking, a deeper analysis revealed that the query protein (GenBank Accession #: NP_717854) had by far the largest number of results from the CDD database (approximately 100). Additionally, the number of trials set in the approximating algorithm was set to 1000, a fairly low initial number to ensure fast

calculation of relevance scores. When this parameter was increased (to 50,000) the additional simulation precision allowed the system to consistently rank the correct COG for the protein. This did impact the performance of the system however as the (observed) time to calculate relevance scores increased significantly (seconds, to minutes). The negative impact of having to increase the trial parameter on the tractability of the relevance score algorithm has illuminated the need for us to explore alternative methods to calculating relevance scores.

The approximating algorithm itself, as previously mentioned in Section 2, is based upon Network Reliability Theory (NRT) [19]. In NRT, the reliability of nodes is affected by the quality and number of paths to that node. Generally speaking, the reliability of a node is high if there exists a high-quality path to the node or there exists multiple paths to the node (convergences). In analogous fashion, data entities in our system can (theoretically) obtain a high relevance score if they are connected to the seed query by a high-quality path or if multiple paths from the seed query converge on the data entity. A limitation of our validation study is that we only evaluate single paths of evidence and not convergences, e.g. there is only a single path of data entities from the seed query protein to a particular COG result. A study which evaluates results from the system where convergences are possible would be extremely interesting but, as it turns out, problematic to perform due to sparse usage of controlled terminologies (such as GO) by biological data sources. For example, it is difficult to determine if two sources attribute the same biochemical function to a particular protein as one may use a synonym. Unless the system can determine that a particular biochemical function is a synonym of another, the paths from the two data sources will not converge. The use of controlled vocabularies by biological data sources is steadily improving however so an evaluation may be possible in the not-too-distant future. Also, calculating relevance scores by simple chaining is proving to be quite valuable in and of itself. For instance, biological domain experts often assess the quality of BLAST results by a single metric (the e-value). The framework of our system allows for the use of additional metrics such as whether or not the result protein has experimental evidence for its function (highly relevant to a biologist) and rank results based on the e-value and the presence or absence of experimental evidence as well, something that is not currently possible with existing interfaces to BLAST.

Uncertainty metrics in our system, while useful, do pose additional challenges for anyone attempting an implementation. For one, it is often difficult to determine the appropriate attributes of data records for Pr and Qr values or the attributes may not be in a usable format. GenBank protein records, for instance, have an "experimental evidence" tag. This is an optional tag so it is often not present. When it is present, using it to determine appropriate Pr values can still be difficult since its value is uncontrolled free text. Also, we have determined the values of our uncertainty metrics through expert opinion and not via quantitative machine learning techniques so it certainly possible that the values may not be generally applicable for other user bases. Determining the values for all uncertainty metrics would be a difficult task however given the heterogeneity and volume of data. Instead, as we mention in future directions, we are planning on doing a sensitivity analysis to determine the robustness of our UII score calculation to variations in uncertainty metrics. Finally, uncertainty metrics add another layer of complexity when creating relationships between data entities in the mediated schema as this affects the calculation of relevance scores. For example, certain biological data sources use algorithmic means to assign a biochemical function to a protein on-the-fly. Protein records themselves however, may store pre-generated references to biochemical functions generated from the same source. If both relationships are represented in the mediated schema this can lead to what is essentially "double-counting" of evidence which can lead to artificial inflation of the relevance score. Unfortunately, in our experience, it is often difficult to account for this as it is not always possible to determine the provenance of any particular data element in biological data sources.

5. Related Work

Not a great deal of work has been done on incorporating uncertainty metrics into data integration systems. Trio [31], and Mystiq [32] are projects whose goal is to build probabilistic database systems but both, however, are focused on creating a centralized database and not an information integration system. Of the data integration systems specifically developed to deal with biological data such as SRS, Kleisli, and Taverna [33-35], none deal with uncertainties in the data.

6. Conclusions & Future Directions

Limitations aside, the results of this study indicate that our work in incorporating data uncertainty metrics into a general-purpose data federated data integration system is foundationally solid and has potential for real-world application in assigning biochemical

function to proteins. While it is true that the results in this study focus on validating results from a single data source, a subsequent study which evaluates results from multiple data sources for functional annotation is currently in progress. As potential future work we plan on implementing and evaluating alternative algorithms for calculating relevance scores to address the tractability issues we have encountered using our current algorithm, which uses simulation techniques and tends to require a large number of trials in order to reach a level of precision sufficient for our needs. We also plan on further evaluating the performance of our current approximating algorithm by carrying out a sensitivity analysis to determine how sensitive the final relevance scores are to various changes in selected uncertainty metrics. Results from this analysis will give us an idea of how "accurate" the uncertainty metrics must be to provide generally correct relevance scores. If the relevance scores turn out to be generally robust then efforts to quantitatively determine the uncertainty metrics, such as using machine learning techniques, may be unnecessary. Additionally, rather than use relevance scores to simply rank results, we would like to use them earlier in the query expansion process to limit the build-up of large result sets. This should also improve the performance of the system overall by enabling the system to follow the most highly relevant paths early in the query process.

7. Acknowledgements

We are thankful for the support of NSF IIS-0513877, NIH NHGRI/NLM R01 HG02288, and NIH NLM T15 LM07442. We would also like to thank Mark Minie, Ph.D for helping us design the COG validation study and to Janos Barbero for developing the graphical user interface to BioMediator.

8. References

[1] D. D. Woods, E. S. Patterson, E. M. Roth, and K. Christofferson, "Can We Ever Escape From Data Overload? A Cognitive Systems Diagnosis," *Cognition, Technology, and Work*, vol. 41, pp. 22-36, 2002.

[2] C. M. Deane, L. Salwinski, I. Xenarios, and D. Eisenberg, "Two Methods for Assessment of the Reliability of High Throughput Observations," *Molecular and Cellular Proteomics*, vol. 1, pp. 349-356, 2002.

[3] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped Blast and Psi-Blast: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, pp. 3389-3402, 1997.

[4] K. Pruitt and D. Maglott, "RefSeq and LocusLink: NCBI gene-centered resources," *Nucleic Acids Research*, vol. 29, pp. 137-140, 2001.

[5] M. Galperin, "The Molecular Biology Database Collection: 2006 update.," *Nucleic Acids Research*, vol. 34, pp. D3-D5, 2006.

[6] L. Stein, "Integrating Biological Databases," *Nature Reviews: Genetics*, vol. 4, pp. 337-45, 2003.

[7] T. G. O. Consortium, "Gene Ontology: tool for the unification of biology.," *Nature Genetics*, vol. 25, pp. 25-29, 2000.

[8] M. Galperin and E. V. Koonin, "'Conserved hypothetical' proteins: prioritization of targets for experimental study," *Nucleic Acids Research*, vol. 32, pp. 5452-5463, 2004.

[9] R. Roberts, "Identifying Protein Function - A Call for Community Action," *PLoS Biology*, vol. 2, pp. e42, 2004.

[10] E. Cadag, B. Louie, P. Myler, and P. Tarczy-Hornoch, "Biomediator Data Integration and Inference for Functional Annotation of Anonymous Sequences," presented at Pacific Symposium on Biocomputing, 2007.

[11] E. Sonnhammer, S. Eddy, and R. Durbin, "Pfam: a comprehensive database of protein domain families based on seed alignments," *Proteins*, vol. 28, pp. 405-20, 1997.

[12] R. Shaker, P. Mork, J. Brockenbrough, L. Donelson, and P. Tarczy-Hornoch, "The BioMediator System as a Tool for Integrating Databases on the Web," presented at Proceedings of the Workshop on Information Integration on the Web, Toronto, ON, 2004.

[13] P. Mork, "Peer Architectures for Knowledge Sharing," in *Computer Science and Engineering*, vol. Doctor of Philosophy. Seattle: University of Washington, 2005, pp. 229.

[14] H. Mei, P. Tarczy-Hornoch, P. Mork, A. Rossini, R. Shaker, and L. Donelson, "Expression Array Annotation Using the BioMediator Biological Data Integration System and the BioConductor Analytic Platform," presented at Proc AMIA Symp, 2003.

[15] L. Donelson, P. Tarczy-Hornoch, P. Mork, C. Dolan, J. Mitchell, M. Barrier, and H. Mei, "The BioMediator System as a Data Integration Tool to Answer Diverse Biologic Queries," presented at Medinfo, 2003.

[16] R. Shaker, P. Mork, M. Barclay, and P. Tarczy-Hornoch, "A Rule Driven Bi-Directional Translation System Remapping Queries and Result Sets Between a Mediated Schema and Heterogeneous Data Sources," *Jour Amer Med Inform Assoc, Fall Symposium Suppl*, pp. 692-696, 2002.

[17] P. Mork, R. Shaker, A. Y. Halevy, and P. Tarczy-Hornoch, "PQL: A Declarative Query Language over Dynamic Biological Schemata," *Jour Amer Med Inform Assoc, Fall Symposium Suppl*, pp. 533-537, 2002.

[18] P. Mork, A. Y. Halevy, and P. Tarczy-Hornoch, "A Model for Data Integration Systems of BioMedical Data Applied to Online Genetic Databases," presented at Proceedings of the American Medical Informatics Annual Fall Symposium, Washington, D.C., 2001.

[19] C. Colbourn, *The Combinatorics of Network Reliability*. New York, NY, USA: Oxford University Press, Inc., 1987.

[20] D. Karger, "A Randomized Fully Polynomial Time Approximation Scheme for the All-Terminal Network Reliability Problem," *SIAM Review*, vol. 43, pp. 499-422, 2001.

[21] R. Tatusov, M. Galperin, D. Natale, and E. V. Koonin, "The COG database: a tool for genome-scale analysis of protein functions and evolution," *Nucleic Acids Research*, vol. 28, pp. 33-36, 2000.

[22] T. G. O. Consortium, "The Gene Ontology (GO) database and informatics resource," *Nucleic Acids Research*, vol. 32, pp. D258-261, 2004.

[23] A. Marchler-Bauer, A. Panchenko, B. Shoemaker, P. Thiessen, L. Geer, and S. Bryant, "CDD: a database of conserved domain alignments with links to domain three-dimensional structure," *Nucleic Acids Research*, vol. 30, pp. 281-283, 2002.

[24] D. Maglott, J. Ostell, K. Pruitt, and T. Tatusova, "Entrez Gene: gene-centered information at NCBI," *Nucleic Acids Research*, vol. 33, pp. D54-58, 2005.

[25] A. Bateman, E. Birney, and E. Sonnhammer, "The Pfam Protein Families Database," *Nucleic Acids Research*, vol. 30, pp. 276-280, 2002.

[26] C. Wu, H. Huang, L. Yeh, and W. Barker, "Protein family classification and functional annotation," *Computational Biology and Chemistry*, vol. 27, pp. 37-47, 2003.

[27] G. Gilliland, T. N. Bhat, H. Weissig, I. Shindyalov, and P. Bourne, "The Protein Data Bank," *Nucleic Acids Research*, vol. 28, pp. 235-242, 2000.

[28] A. Barioch, R. Apweiler, C. Wu, W. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. Martin, D. Natale, C. O'Donovan, N. Redaschi, and L. Yeh, "The Universal Protein Resource (UniProt)," *Nucleic Acids Research*, vol. 33, pp. D154-9, 2005.

[29] J. Schultz, F. Milpetz, P. Bork, and C. P. P. & Ponting, "SMART, a simple modular architecture research tool: Identification of signaling domains.," *PNAS*, vol. 95, pp. 5857-5864, 1998.

[30] B. Boeckmann, A. Barioch, R. Apweiler, and M. Schneider, "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003," *Nucleic Acids Research*, vol. 31, pp. 365-370, 2003.

[31] J. Widom, "Trio: a system for integrated management of data, accuracy, and lineage.," presented at Proceedings of the Second Biennial Conference on Innovative Data Systems Research (CIDR), 2005.

[32] J. Boulos, N. Dalvi, B. Mandhani, S. Mathur, C. Re, and D. Suciu, "MYSTIQ: A system for finding more answers by using probabilities," presented at SIGMOD, Baltimore, Maryland, USA, 2005.

[33] T. Oinn, M. Addis, J. Ferris, D. Marvin, M. Senger, M. Greenwood, T. Carver, K. Glover, M. Pocock, A. Wipat, and P. Li, "Taverna: a tool for the composition and enactment of bioinformatics workflows," *Bioinformatics*, vol. 20, pp. 3045-3054, 2004.

[34] S. Y. Chung and L. Wong, "Kleisli: a new tools for data integration in biology," *Trends in Biotechnology*, vol. 17, pp. 351-355, 1999.

[35] T. Etzold, A. Ulyanov, and P. Argos, "SRS: information retrieval for molecular biology data banks," *Methods Enzymology*, vol. 266, pp. 114-28, 1996.